STATISTICS NOTES

Dept of Psychiatry, KIMSRF

Sl. No	Content	Page no.
1	Basic definitions in Statistics	1
2	Scales for data representation	3
3	Sampling	4
4	Sample size	5
5	Measures of central tendency	8
6	Reliability	9
7	Validity	10
8	Normal distribution	11
9	Types of error	12
10	Parametric tests	14
11	Student t-test	14
12	ANOVA	15
13	Non Parametric tests	16
14	Chi Square test	17

Statistics: It is a broad subject, with applications in a vast number of different fields. Generally speaking it is the methodology of collecting, analysing, interpreting and drawing conclusions from information. It provides methods for

- Design: planning and carrying out research studies.
- Description: summarising and exploring data.
- Inference: making predictions and generalising about phenomena represented by the data.

Population and Sample

- Two basic concepts
- Population: it is the set of individual persons or objects in which an investigator is primarily interested during his research problem.
- Sample: it is that part of the population from which information is collected.

Descriptive and Inferential statistics

- Two major types of statistics
- Descriptive statistics: it consists of methodology for organising and summarising information. It provides us the general overview of the data. It includes construction of graphs, charts, tables and calculation of measures like averages, measures of variation and percentiles.

SHOUT
SHARE
IT

- Inferential statistics: it consists of methods for drawing and measuring the reliability of conclusions about population based on the information obtained from a sample of the population. It includes point estimation, interval estimation and hypothesis testing. Ex are frequency table, class interval, bar graph etc.
- Both are inter-related. It is always necessary to use methods of descriptive statistics to organise and summarise the information before methods of inferential statistics can be used. Ex are chi square test, t test, ANOVA etc.
- Preliminary descriptive statistics reveals the choice of the appropriate inferential method to be used.

Variable: it is any characteristic that varies from one individual member of the population to another. Ex are height, weight, age, marital status etc.

Types of variables

- Quantitative variable: they yield numerical information. Ex is age, height etc. it is further divided into Discrete and Continuous.
- Discrete variable: it can take only specified values- often integers. No intermediate numbers are possible. Ex is number of children, students taking exam etc.
- Continuous variable: data are not restricted on taking any certain specified value. Fractional values are possible. Ex is height, weight etc. Weight can be measured accurately to the tenth of a gram.
- Qualitative variable: they yield non numerical information. Ex is marital status, sex etc.

Tables –It is a simplest means of summarising a set of observations and can be used to represent all types of data

- Frequency distribution
    - The number of observations that fall into a particular class of the qualitative variable is called as frequency. A table listing all the classes and their frequencies as called a frequency distribution.
    - For Nominal and Ordinal data, a frequency distribution consists of a set of classes or categories with the numerical count that corresponds to each one.
    - To display Interval or Ratio data, data must be broken down into a series of distinct, non overlapping intervals called as class intervals (CI).
    - If there are too many class intervals, the summary is not too much of an improvement over raw data. If there are too few class intervals then a great deal of information is lost.
    - Usually class intervals are constructed so that all have equal width; this facilitates comparison between the classes.
    - Once the upper and lower limit for each class interval is selected, the number of values that fall within that pair of limit is counted and the result is arranged as a table.
- Relative frequency
    - This gives the proportion of values that fall into a given interval in frequency distribution.

- It is calculated by dividing the number of values within that interval by total number of values in that table.
- It can be useful in comparing sets of data that contain unequal numbers of observations.

SBP (mmHg)	Frequency	Relative frequency
100-109	5	10
110-119	15	30
120-129	21	42
130-139	6	12
140-149	3	6
Total	50	100

Scales for data representation

- The scale gives certain structure to the variable and also defines the meaning of the variable.
- There are four types of scale Nominal, Ordinal, Interval and Ratio scales
- Nominal:
  - Simplest type of data.
  - Values fall in unordered categories or classes.
  - Nominal refers to the fact the categories are merely names.
  - Used to represent Qualitative data.
  - Ex include gender- male & female, blood groups- A, B, AB & O
  - If the data has only two distinct values then it is called as Binary/ Dichotomous.
- Ordinal:
  - If the same nominal can be put into an order then it is called as Ordinal data.
  - Used to represent Qualitative data.
  - Ex include depression as mild, moderate & severe.
  - Here a natural order exists among the groups, but the difference between the groups is not necessarily equal.
- Interval:
  - Here the data can be placed in a meaningful order and there is equal difference between the groups.
  - Ratio of the measurements cannot be done, that is there is no absolute zero.
  - Used to compare Quantitative data.
  - Ex include temperature measured in centigrade, time (even though 0000hrs exist, there is nothing like no time)
- Ratio:
  - Here there is comparable difference between the variables as well as there is absolute zero.
  - Ratio of the measurements also can be done.
  - Used to measure Quantitative data.
  - Ex include temperature measured in Kelvin.

Properties of data-

- Central tendency- Mean, Median & Mode
- Variation- Range, Variance, Standard Deviation & Coefficient of Variation
- Shape- Skewness & Kurtosis

Scale type	Permissible statistics	
Nominal	Mode	Chi-Square test
Ordinal	Mode/ Median	
Interval	Mean, Standard Deviation	Correlation, Regression, t-Test, ANOVA
Ratio		

Sampling

- Sampling is a process used in statistical analysis in which a predetermined number of observations will be taken from a larger population.
- A population includes all people/ items with the characteristic one wishes to understand. Because there is very rarely enough time or money to gather information from everyone/ everything in a population, the goal becomes finding a representative sample of that population.
- Probability sampling- every unit in the population has a chance of being selected in the sample. This makes it possible to produce unbiased estimates of population. It includes simple, Systemic, Stratified and Cluster sampling.
- Non probability sampling- here some elements of the population have no chance of selection. It involves the selection of elements based on assumptions regarding the population of interest, which forms the criteria for selection. It includes Quota, Purposive and Convenience sampling.
- Types of sampling techniques
    1. Simple random sampling
        - Each element in the population has an equal probability of selection and each combination of elements has an equal probability of selection.
        - Numbers are drawn out by lottery technique.
    2. Systematic random sampling
        - Relies on arranging the study population according to some ordering scheme and then selecting elements at regular intervals.
        - It involves a random start and then proceeds with selection of every kth element from then on.
        - It is less accurate than simple random sampling.
    3. Stratified sampling
        - When a population has a number of distinct categories, the frame can be organised by these categories into separate 'starta'.
        - Each stratum is then sampled as an independent sub population, out of which individual elements can be randomly selected.
    4. Cluster sampling
        - Sometimes it becomes impossible/ impractical to create a sampling frame of a target population, and/ or the target population is widely dispersed

geographically, making data collection costs relatively high. In such cluster sampling would be ideal.

- A cluster is an aggregate or intact grouping of population. This type of sampling involves random selection of items from clusters.
- Clusters may be space based, such as naturally occurring geographical/ physical units; organisation based like districts, schools, grades etc.
- Heterogeneity of the cluster is central to a good cluster sample design. Ideally, within cluster difference should be high and between cluster difference should be low.
- It is commonly implemented in multistage sampling.

5. Quota sampling
    - This is similar to stratified sampling, but instead of randomisation quota is used in selecting the sample.
6. Purposive sampling
    - Sample is selected based on the purpose of the research.
    - Here the researcher believes that the sample selected serves the purpose of the research.
    - It has big sampling errors and carry misleading conclusions.
7. Convenience sampling
    - Sample is selected because it is easy to assess/ collect them.
    - No reason tied to the purpose of research.
8. Snowball sampling
    - Sociometric sampling technique generally used to study the small group.
    - All the persons in a group identify their friends who in turn identify their friends and colleagues, until the informal relationships coverage into some type of a definite social pattern.
    - It is just like snowball increasing in size as it goes on rolling in ice field.

What do you understand by sample size in a research study? What are the statistical methods used to determine the sample size? (5+5)

Introduction- Planning a study with a sample size calculation is an important aspect. It is not possible to study the whole population in any study. Thus a set of participants, who represent the population, are taken and study is conducted on them. This represents a sample; the results thus obtained on sample are extrapolated to the whole population.

Definitions-

Population- it consists of all elements- individuals, items or objects- whose characteristics are being studied. The population that is being studied is also called target population.

Sample- it is a subset of a population on which the statistical studies are made in order to draw conclusions relative to the population.

Sample size- it is the number of individuals or elements belonging to a sample.

Attributes of a sample-

- Every individual in the population should have an equal chance to be included in the sample.
- Choice of one participant should not affect the chance of another's selection.

Problems associated with not estimating sample size - Randomly taking a sample size gives rise to problems as-

1. Ethical problems- too many patients are given inferior treatment, which is not ethically correct.
2. Scientific problems- negative research require the repetition of the research to firmly prove the point.
3. Financial problems- extra costs are involved in too small and too large studies.

Sample size depends on the following factors-

- Level of significance ("p" value):- Prior to starting of the study an acceptable value is set. Usually it is accepted at $p < 0.05$, it means we are ready to accept 5% results by chance factor. In another words 5% of false positive is accepted.
- Power of the study (β error):- Here we fail to detect a difference when there is actually a difference. This is called false negative. This must be decided before the study starts. Usually most studies accept a power of 80%, it means that in one in five (i.e is 20%) times we will miss a real difference. For some important studies it is set at 90%, to reduce the possibility of false negatives to 10%.
- Expected effect size:- It is the difference between the value of variable in the control group and that in the test group. It can be calculated from the previous reported or preclinical studies.
- Underlying event rate in the population:- Estimated from the previous studies.
- Standard deviation in the population:- It is the measure of dispersion or variability in the data. It is easy to understand that we require a smaller sample if the population is more homogenous and therefore has a smaller variance and standard deviation.

Methods of calculating sample size-

1. *Using Standard Error of Means (SEM)-*

   Mean should be atleast 1.96 or approx 2 SEMs distant from 0 to obtain statistical significance.

   Assuming: mean = 2 SEM,
   Then: mean/ SEM = 2,
   Then: mean/ SD/ $\sqrt{n}$ = 2,
   Then: $\sqrt{n}$ = 2 * SD / mean,
   Then: $n = 4 * (SD/ mean)^2$.

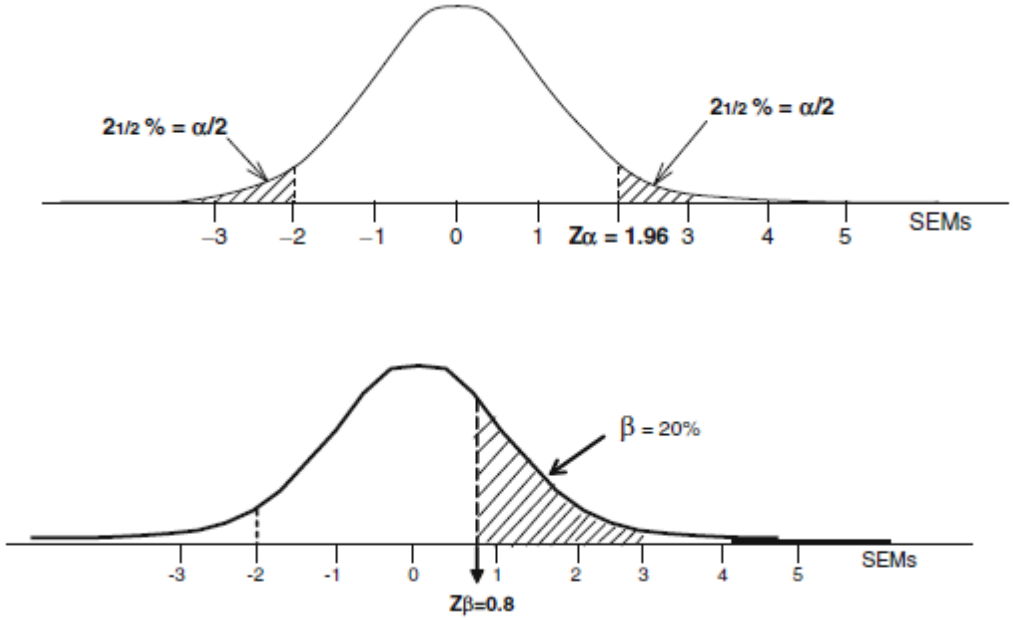
2. *Using power index method-*

To calculate the sample size, we need to know statistical power, mean difference and variance in the data estimated as SD or SEM.

The relationship between the elements can expressed with, $(z_\alpha+z_\beta)^2$ = Power index, being the important element as: n = (SD / mean)^2 $(z_\alpha+z_\beta)^2$

Z_α means a place on Z line, if α is defined as 5%, then right from this place on the Z line AUC (Area Under Curve) = 5%.so this place must be 1.96 SEMs distant from 0. So Z_α = 1.96 or approx 2.0.

Z_β, where β is defined 20%, it is right form this place the AUC= 20% of the total AUC. This place is approx 0.6 SEMs distant from 0. So Z_β = approx 0.8.



So power index can be calculated as, power index = $(z_\alpha+z_\beta)^2$ = $(2.0+0.8)^2$ = 7.8. So for continuous variable with α = 5% and power= 1-β = 80% the required sample size is-

N = 7.8 * (SD / mean)^2.

3. *Calculation for parallel group studies*

Parallel group studies including 2 groups larger sample sizes are required. Each group produces its own mean and standard deviation (SD).

The pooled SD = √(SD_{group1}^2 + SD_{group2}^2)

The equation for sample size is given by:

n= 2 $(z_\alpha+z_\beta)^2$ (pooled SD/ mean difference)^2.

4. *For studies with proportions*

   The equations for continuous data and proportions are very similar-

   Continuous data: n = power index * (SD / mean)^2.
   Proportions: n = power index * (SD / proportion)^2.

   For parallel group studies with two proportions again SDs are pooled-

   Pooled SDs = $\sqrt{(SD_{group1}^2 + SD_{group2}^2)}$

   Then the equation becomes: $n = 2(z_\alpha+z_\beta)^2 * p_1(1-p_1) + p_2(1-p_2)/ (p_1-p_2)^2$, where p_1 and p_2 are proportions of groups 1 and 2 respectively. Even though the equation looks different it is similar to that of the continuous data.

Factors affecting the sample size-

1. The sample size increases as variance (σ^2) increases.
2. The sample size increases as the significance level is made smaller (α decrease).
3. The sample size increases as the required power increases (1-β increases).
4. The sample size decreases as the absolute value of the distance between the null and alternative means ($|\mu_0- \mu_1|$) increases.

References-

1. Kadam P & Bhalerao S. Sample size calculation. Int J Ayurveda Res. 2010 Jan-Mar; 1(1): 55-57.
2. Statistics applied to clinical studies, 5th edition by Ton J.Cleophas & Aeilko H.Zwinderman. Springer.

Measures of Central tendency

- In any distribution, majority of the observations pile up, or cluster around in a particular region. This is referred to as the central tendency of the distribution.
- So it is a statistical measure to determine a single score that defines the centre of the distribution.
- It makes the large amount of information easily understandable.
- There are three types of central tendency- Mean, Median & Mode
- Mean and Median can be applied only to Quantitative data, whereas Mode can be used with either Quantitative or Qualitative data.
- Mean-
    - Most frequently used measure of central tendency
    - It is the sum total of all the observations divided by total number of observations.
    - It is stable average based on all observations.
    - Calculated for Quantitative data, measured in interval/ ratio scale
    - It is affected by extreme values (outliers), because it is extremely sensitive to unusual values.

    - Ex- set of data 5, 2, 8, 6, 7, 3, 11, 6, mean would be (5+2+8+6+7+3+11+6/8) that would be 48/8= 6
- Median-
    - Scores that divides a distribution exactly into half. It is defined as the 50^{th} percentile of a set of measurements.
    - Exactly 50% of the scores fall above the median and 50% below.
    - To find the median, the data has to be arranged in ascending/ descending order and find the midpoint.
    - If the total number of score is a even number, then average of two data is taken.
    - It can be calculated for quantitative data with extreme values (outliers). That is when the data is not symmetrically distributed.
    - If a set of data contain 'n' (odd) observations, median is the middle observation or [(n+1)/2]th value. If data has even observations, then median is taken as average of two middle observations.
    - Ex- set of data 5, 2, 8, 6, 7, 3, 11, 6, median would be 2, 3, 5, 6, 6, 7, 8, 9, 11 average of 4^{th} & 5^{th} value, that is 6+6/2=6
- Mode-
    - It is the most frequently occurring observation, i.e. around this value the observations tend to be most heavily concentrated.
    - Not affected by extreme values (outliers)
    - Some distributions could have several modes, while some could have no mode.
    - Ex- set of data 5, 2, 8, 6, 7, 3, 11, 6; mode would be the most commonly occurring value that is 6.
    - If the data has no value occurring more than once, then the data does not have mode.
- Which measure to choose-
    - Mean- when the variable is quantitative with symmetrical distribution.
    - Median- in case of quantitative variable with skewed distribution.
    - Mode- used for qualitative variable.

Reliability

- Definition: Reliability refers to the ability of a measurement instrument to produce the same result on repeated measurement.
- Types of reliability:
    1. Scorer/ Inter-rater reliability:
        - It refers to ability of measurement instrument to produce the same result when administered by two different raters.
        - It is the probability that 2 raters (i) will give the same score to given answer, (ii) rate a given behaviour in the same way, (iii) add up the score properly.
        - Scorer reliability should be near perfect.
    2. Test-Retest reliability:
        - It assesses the ability of a measurement to arrive at the same result for the same subject on repeated administrations.

        - The interval between the test and retest should be long enough to ensure that the person's responses are based on his/ her current condition rather than the memory of responses in the first test administration.
        - If there is too long, there is a risk that the person's condition may have changed.
    3. Parallel form reliability:
        - It refers to the degree to which two equivalent versions of a test give the same result.
        - This type of reliability is usually used when a test cannot be exactly repeated.
    4. Split half reliability:
        - If a test cannot be repeated or if there is no parallel form, a test can be split in two and these two halves are correlated with each other. For e.x odd vs even items.
        - There is a mathematical formula for computing the mean of all possible split halves.
    5. Internal consistency:
        - The degree to which one test item correlates with all other test items.
        - It is denoted by α co-efficient.
        - It should not drop below 0.7.
- Psychometric tests aim to measure a real quantity. Real quantity is true score (t), score obtained on test is observed score (x). As no test is perfect, there is error (e).
- Aim is to reduce 'e' to minimum to make as reliable as possible.
- In practice, when the test is repeated, each occasion will give different score, i.e. observed score will cluster around true score. Like the distribution of any variable, the distribution of observed score would have mean and SD.
- If reliability of the test is perfect, real score= observed score (t=e).

Validity

- Definition: the degree to which a test measures what it is supposed to measure is known as validity.
- Types of validity:
    1. Face validity:
        - It refers to whether test seems sensible to the person completing it, i.e. does it appear to measure what it is meant to be measuring.
    2. Content validity:
        - It refers to the degree to which the test measures all the aspects of the item that is being assessed.
        - For example, test for depression should have questions asking depressive symptoms.
    3. Concurrent validity:
        - It reveals that at a given point of time, high scorers on a test may be more likely than low scorers.

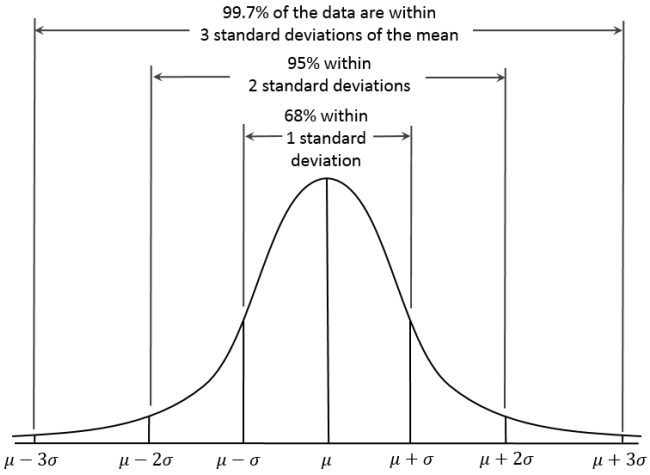
        - To determine test's concurrent validity, external measures are obtained at the same time that the test is given to the sample of subjects.
        - For example, correlation between HAMD and MADRS in which concurrent validity of MADRS is checked with HAMD, already established instrument for depression.
    4. Predictive validity:
        - It refers to degree to which a test predicts whether some criterion is achieved in future.
        - Here the external criterion would have to be obtained a number of years down the road for the test to have the predictive validity.
        - For example. Whether a child IQ test predicts
    5. Construct validity:
        - It refers to whether a test measures some specified hypothetical construct.
        - If a test is measuring one construct, there should not be cluster of items that seem to be measuring different things.
    6. Factorial validity:
        - If a test breaks down into various sub factors then the number and nature of these factors should remain stable across time and different subject populations.
    7. Incremental validity:
        - It refers whether the test results improve decision making.
        - For example, whether knowledge of neuropsychological test results improves the detection of brain injury.

Normal distribution

- Also called as Bell shaped curve or Gaussian distribution, after the well known mathematician Karl Freidrich Gauss.
- It is the most common and widely used continuous distribution
- Bell shaped curve can be obtained by compiling data into a frequency table and graphing it in a histogram.
- Normal distribution is easy to work with mathematically. In many practical cases, the methods developed using normal theory work well even when the distribution is nearly normally distributed.
- Standard normal distribution (Z distribution), is used to find probabilities and percentiles for regular normal distributions. It serves as a standard by which all other normal distributions are measured.
- It is a normal distribution with mean "0" and standard deviation of "1".
- Properties of the standard normal curve
    - Its shape is symmetric.
    - Area of the curve is greatest in the middle, where there hump and thins out towards the tails.
    - It has mean denoted by μ (mu) and standard deviation denoted by б (sigma).
    - Total area under the curve is equal to 1.

- Property of the curve is mean= median= mode, this is because the shape of the data is symmetrical with one peak.
- Because of bell shape, probabilities for the normal distribution follow the empirical rule. About 68% of values lie within 1 SD of mean, 95% of values lie within 2 SD of mean and almost its entire values (about 99.7%) lie within 3 SDs of the mean.
- There is a strong correlation between the size of a sample n and the extent to which a sampling distribution approaches the normal distribution.

- Importance in statistics- All the inferential statistics like correlation, regression, t test and ANOVA are based on the assumptions that the data follows a normal distribution. It is assumed that the samples taken are normally distributed.
- The once which do not use these properties are called Non parametric tests.
- Tests to check normal distribution- it is important to check whether the given data follows a normal distribution or not. If it does not then parametric test cannot be used.
  - By measuring central tendency, skewness, kurtosis.
  - Graphical representation of the data.
  - Tests like Kolmogorov-Smirnov test, Shapiro-Wilk test etc.
- If data is not normally distributed, the data can be transferred so that parametric test can be applied. It can be done by log transformation or square root transformation.



Types of error

- Aim of doing a study is to check whether the data agree with certain predictions. These predictions are called hypothesis.
- Hypothesis arises from the theory that drives the research. These are formulated before collecting the data.
- Hypothesis testing will help to find if variation between two sample distributions can just be explained through random chance or not. Before concluding that two distributions vary in a meaningful way, precautions must be taken to see that the differences are not just through random chance.
- There are two types of hypothesis- Null (H_0) and Alternative hypothesis (H_1).

- Null hypothesis is usually a statement that the parameter has value corresponding to, in some sense, no effect.
- Alternative hypothesis is a hypothesis which contradicts the null hypothesis.
- Significance test is a way of statistically testing hypothesis by comparing the data values.
- Significance test analyses the strength of sample evidence against the null hypothesis. The test is conducted to investigate whether the data contradicts the null hypothesis, suggesting alternative hypothesis is true.
- *p*-Value is the probability, if H_0 were true, that the test statistic would fall in this collection of values. The smaller the *p*-value, the more strongly the data contradicts H_0. When p-value ≤ 0.05, data sufficiently contradicts H_0.
- **Type α/ I error**- Rejecting **true** null hypothesis.
    - Leading to conclusion that difference is significant, when in fact there is no real difference.
    - In simple terms, it is asserting something that is absent. It is False Positive.
    - For example, a study indicating that a particular treatment cures a disease when in fact it does not.
    - It is popularly known as ***p*-value**. Maximum *p*-value allowed is called as **level of significance**. Being serious p-value is kept low, mostly less than 5% or **p<0.05**.
    - It means to say that there is only 5 in 100 chance that the variation we are seeing is due to chance.
- **Type β/ II error**- Accepting **false** null hypothesis.
    - Leading to conclusion that difference is not significant, when in fact there is real difference.
    - It is also called as False Negative.
    - For example, a study indicating that a new treatment modality fails to work, when in fact it does work.
    - It is also called as **Power** of the test & indicates sensitivity of the test.
    - Type II error can be decreased by ensuring enough power.
- All statistical hypothesis tests have a probability of making type I and II errors. These error rates are traded off against each other. For a given test, the effort to reduce one type of error generally results in increasing the other type of error.
- For a given test, one way to reduce the error is to increase the sample size where ever it is feasible.
- It is not possible to reduce both type I & II, So α error is fixed at a tolerable limit & β error is minimized by ↑ sample size.

		Truth	
	Hypothesis Testing	**The Null Hypothesis Is True**	**The Alternative Hypothesis Is True**
Research	**The Null Hypothesis Is True**	Accurate	Type II Error
	The Alternative Hypothesis Is True	Type I Error	Accurate

Parametric tests

- A parametric test is one that makes assumptions about the parameters (defining properties) of the population distribution from which one's data are drawn.
- These tests assume that data is normally distributed and rely on group means.
- Most well known elementary statistical methods are parametric.
- These tests assume more about a given population. When the assumptions are correct, they produce more accurate and precise estimates.
- When the assumptions are not correct these tests have a greater chance of failing, and for this reason are not robust statistical methods.
- Parametric formulae are often simpler to write down and faster to compute, and for this reason they lack robustness.
- Examples of Parametric tests are t-tests and ANOVA.
- Advantages over Non parametric tests
    1. Statistical power- parametric test have more statistical power than non parametric tests. Thus more likely to detect a significant effect when one really exists.
    2. Parametric perform well when the spread of each group is different- this is because even though non parametric tests don't assume that data follow normal distribution, they have other assumptions like the data for all groups must have the same spread.
    3. Parametric tests can perform well even with skewed and non normal distributions, provided the data satisfies some guidelines about sample size.

Parametric test	*Sample size guidelines for non normal distribution*
1-Sample t test	> 20
2-Sample t test	Each group > 15
One way ANOVA	If 2 -9 groups, then each group > 15 If 10 -12 groups, each group > 20

Student t-test

- Developed by WS Gosset, a chemist working for the Guinness brewery in Dublin, Ireland ("Student" was his pen name).
- It can be used to determine if two sets of data are significantly different from each other.
- A t-test is any statistical hypothesis test in which the test statistic follows a Student's t distribution if the null hypothesis is supported.
- The t-distribution is similar to standard normal distribution, but it's shorter and flatter than the standard normal distribution (Z distribution). But as the sample size increases (degrees of freedom) the t distribution approaches the standard normal distribution.
- Degree of freedom is always a simple function of sample size that is (n-1).
- Particular advantage of the t-test is it does not require any knowledge of the population standard deviation. So it can be used to test hypothesis of a completely unknown population, and the only available information about the sample comes from the sample.
- All that is required for a hypothesis test with t-test is a sample and a reasonable hypothesis about the population mean.
- Types of Student t-test

1. One sample t-test: only one group and typically used to compare a sample mean to a known population mean. E.g. a group of schizophrenic patients were assessed on cognitive tests and the group is compared with the population.
2. Paired/ Dependent sample t-test: there are two groups and two means which are related to each other. Like two scores for each person, or when there are matched scores. E.g. a sample is tested to see the metabolic side effects before and after treatment. Before and after sample are compared to see any difference.
3. Unpaired/ Independent/ two sample t-test: there are two means, two groups and not related to each other. Used to compare means from independent groups. E.g. two sample of patients one given a placebo and another newer anti psychotic both were compared to see the utility of newer antipsychotic.

- Assumptions:
    1. Data must come from a population that follows a normal distribution.
    2. For two sample t-test, the two populations must have equal variances. If variances are not equal then Wlech's t-test is used.
    3. Each score must be independent of all other scores.

Analysis of Variance (ANOVA)

- It was developed by R.A. Fisher.
- It is a collection of statistical models used to analyze the differences between group means and their associated procedures.
- In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are equal, and therefore generalizes the t-test to more than two groups.
- Doing multiple two-sample t-tests would result in increased chance of committing a type I error, which is avoided by doing ANOVA.
- Characteristics of ANOVA are:
    1. As exploratory data analysis, an ANOVA is an organization of additive data decomposition, and its sums of squares indicate the variance of each component of the decomposition.
    2. Comparisons of mean squares, along with F-tests ... allow testing of a nested sequence of models.
    3. It is used to develop and confirm an explanation for the observed data.
    4. It is computationally elegant and relatively robust.
    5. ANOVA provides industrial strength (multiple sample comparison) statistical analysis.
    6. It has been adapted to the analysis of a variety of experimental designs.

- **Three classes of models:**
    1. **Fixed-effects** model assumes that the data come from normal populations which differ in their means.
    2. **Random-effects** models assume that the data describe a hierarchy of different populations whose differences are constrained by the hierarchy.
    3. **Mixed models** describe situations where both fixed and random effects are present.

- **Types of ANOVA** depending on the number of treatments and the way they are applied to the subjects in the experiment:
    1. **One-way ANOVA** is used to test for differences among >=3 independent groups. E.g. Group A is given vodka, Group B is given gin, and Group C is given a placebo. All groups are then tested with a memory task.

2. **One-way ANOVA for repeated measures** is used when the subjects are dependent groups; this means that the same subjects are used for each treatment. E.g. Group A is given vodka and tested on a memory task. The same group is allowed a rest period of five days and then the experiment is repeated with gin. Again, the procedure is repeated using a placebo.
3. **2×2 ANOVA**, the most common type of factorial analysis of variance, is used when the experimenter wants to study the effects of >= 2 treatment variables. Factorial ANOVA can also be 2×2×2, 3×3, etc. but higher numbers of factors is rarely done because the calculations are lengthy and the results are hard to interpret. E.g. In an experiment testing the effects of expectation of vodka and the actual receiving of vodka, subjects are randomly assigned to four groups: 1) expect vodka-receive vodka, 2) expect vodka-receive placebo, 3) expect placebo-receive vodka, and 4) expect placebo-receive placebo (the last group is used as the control group). Each group is then tested on a memory task. The advantage of this design is that multiple variables can be tested at the same time instead of running two different experiments.

- Assumptions of ANOVA

1. Independence of cases - this is the design requirement.
2. Normality - the distributions in each of the groups are normal.
3. Scale of measurement - dependent variable is interval or ratio.
4. Homogenity of variances - variance of data in groups should be the same.

Non Parametric tests

- Many of the tests that are used for analyses of data have a presumption that data must have a normal distribution. When the data does not meet normal distribution, Non Parametric tests are used.
- These are those tests that don't have any presumption about data.
- These work with Median, which is a much more flexible statistic because it is not affected by outliers.
- Advantages of non parametric tests
    1. Area of study is better represented by median. That is median better represents the centre of distribution.
    2. When the sample size is small. With small sample size it is not possible to ascertain the distribution of data, so the parametric tests lack sufficient power to provide meaningful results.
    3. Presence of outliers. Parametric test can only assess continuous data and the results can be significantly affected by outliers.

	Parametric test	**Non Parametric test**
Testing a mean	One sample t test	Sign test
Comparison of means of 2 unrelated groups	Independent sample t test	Mann Whitney U test
Comparison of means of 2 related samples	Paired t test	Wilcoxon Signed rank test
Comparison of means of > 2 unrelated samples	ANOVA	Kruskal Wallis test
Comparison of means of > 2	Repeated measures of ANOVA	Friedman's test

related samples		
Assessing the relationship between 2 quantitative variables	Pearson's correlation	Spearmen's correlation

Chi Square test

- It is developed by Karl Pearson.
- It is a non parametric test of significance not based on any assumption/ distributions of any variable.
- It is applied to sets of categorical data to evaluate hoe likely it is that any observed difference between the sets arose by chance. It is based on frequencies and not on parameters like mean and SD.
- It is statistical hypothesis test in which the sampling distribution of the test statistic is a chi square distribution when null hypothesis is true.
- Degrees of freedom (df): it denotes the extent of independence/ freedom enjoyed by a given set of observed frequencies.
- If df > 2→ distribution is bell shaped, df = 2 → distribution is 'L' shaped with maxim ordinate at zero and df < 2 → distribution is 'L' shaped with maxim ordinate at the origin.
- Uses: there are three important uses of chi square test
    1. Test of association
    2. Test of goodness of fit of distribution
    3. Test of proportions
- Test of association:
    - Most important application of the test.
    - It measures the probability of association between two discrete attributes. For example, smoking and cancer, treatment and outcome of disease or vaccination and smoking etc.
    - There are two possibilities- either they influence or affect each other or they do not. In other words, they are either independent of each other or they are dependent on each other.
- Test of goodness of fit of distribution
    - Test helps to find whether the observed frequency distribution fits in a hypothetical/ theoretical or assumed distribution of qualitative data.
    - Assumption of no difference between the observed and hypothetical distribution is made (null hypothesis). Frequency assumed to be equal in each case or either case.
- Test of proportions
    - To find significance of difference in two or more than two proportions.
    - For example: incidence of diabetes in 20 non obese with that 20 obese persons or number of diabetics and non diabetics in groups weighing 40-50 kg, 50-60 kg and 60-70 kg.
- Assumptions for applying chi square test
    1. Random sample
    2. Data must be in the form frequencies

3. Frequency data must have a precise numerical value and organised into groups
4. All the in the sample must be independent
5. Lowest expected frequency should not be less than 5

- Limitations of chi square test
    1. Can be applied in only on fourfold table
    2. Will not give result if the expected value in any cell is less then 5
    3. If the sample total is less than 50, need to be interpreted in caution
    4. Test only tells about presence or absence of association between two events and not strength of association
    5. Tells only about probability of occurrence and does not indicate cause and effect relationship

How to use statistics in thesis?

- The use of statistics begins from even before the study is started. It starts from designing the study like what type of study (retrospective/ prospective, case control/ cohort, cross sectional/ follow up), how to collect the sample (simple/ stratified random, convenient) etc.
- Sample size calculation is the most important before the study is actually started. There are various methods in which it can be calculated. Various softwares are available to calculate the same.
- Once the sample size is obtained, the data is collected and has to be transferred onto a MS Excel sheet or a SPSS data sheet.
- SPSS data sheet helps in analysing the data easily. Even a MS Excel data sheet can be converted into SPSS data sheet without much problem using SPSS software.
- The data thus obtained is analysed using descriptive and inferential statistics.